REMARKS

Claims 1, 3-6, and 8-14 are pending in the present application. Claims 1, 11, and 12 are amended. Claims 2 and 7 were previously canceled in a correspondence filed on December 21, 2004. Reconsideration of the claims is respectfully requested.

I. 35 U.S.C. § 103, Obviousness, Claims 1, 3-6, 8, and 10-13

The Examiner has rejected claims 1, 3-6, 8, and 10-13 under 35 U.S.C. § 103 as being unpatentable over Shriberg et al., ELIZABETH SHRIBERG ET AL., SPEECH COMMUNICATION 32 (2000) 127-154 ("*Shriberg*") in view of Yeldener et al., U.S. Patent No. 5,774,837 ("*Yeldener*"). This rejection is respectfully traversed.

The Examiner bears the burden of establishing a *prima facie* case of obviousness based on the prior art when rejecting claims under 35 U.S.C. § 103. *In re Fritch*, 972 F.2d 1260, 23 U.S.P.Q.2d 1780 (Fed. Cir. 1992). For an invention to be prima facie obvious, the prior art must teach or suggest all claim limitations. *In re Royka*, 490 F.2d 981, 180 USPQ 580 (CCPA 1974). The Examiner has not met this burden because all of the features of these claims are not found in the cited references as believed by the Examiner. Therefore, a combination of *Shriberg* and *Yeldener* would not reach the presently claimed invention in these claims.

Amended independent claim 1 of the present invention, which is representative of amended independent claims 11 and 12, reads as follows:

> 1. A method for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of:
>
> determining a fundamental frequency for the digitized audio stream;
>
> detecting changes of the fundamental frequency in the audio stream, wherein detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value, and wherein the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice, and wherein the voicedness of the fundamental frequency estimates higher than the threshold value equals voice;

determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency;

extracting and combining a plurality of prosodic features in a neighborhood of the candidate boundaries; and

determining boundaries for the semantic or syntactic units depending only on the combined plurality of prosodic features.

With regard to claim 1, the Examiner stated:

> Regarding claims 1 and 11-12, Shriberg et al. disclose a method, a computer usable medium having computer readable program code, and a digital audio processing system for the segmentation of an audio stream into semantic or syntactic units wherein the audio stream is provided in a digitized format, comprising the steps of: determining a fundamental frequency for the digitized audio stream (Section 2.1.2.3 on page 133); detecting changes of the fundamental frequency in the audio stream (pages 134-135, refer to figure 4); determining candidate boundaries for the semantic or syntactic units depending on the detected changes of the fundamental frequency (pages 134-135); extracting and combining a plurality of prosodic features in the neighborhood of the candidate boundaries (section 2.1.1 on page 130 and section 2.1.4 on page 137); and determining boundaries for the semantic or syntactic units depending on the at least one prosodic feature (pages 134-135, F0 is a prosodic feature).
>
> **Shriberg et al. fail to specifically disclose the step of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value.** However, Yeldener teaches the step of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value (col. 15, ln. 1 to col. 16, ln. 14 and/or col. 14, ln. 4-55, the goal is to use 0 and 1 to represent for unvoiced and voice portions, respectively).
>
> Since Shriberg et al. and Yeldener et al. are analogous art because they are from the same filed of endeavors, it would have been obvious to one of ordinary skill in the art at the time of invention to modify Shriberg et al. by incorporating the teaching of Yeldener et al. in order to enable the system to pay more coding emphasis on the voice portion than unvoiced portion to reduce processing time and increase transmission rate. (emphasis added).

Final Office Action, dated April 28, 2005, Pages 3 and 4.

Applicants agree with the Examiner that *Shriberg* does not teach or suggest detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value as recited in independent claim 1. Consequently, because *Shriberg* does not teach or suggest the above recited claim 1 feature, *Shriberg* cannot teach or suggest that the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voicc as further recited in amended claim 1.

Yeldener does not curc the above stated deficiency of *Shriberg*. The Examiner cites *Yeldener* as teaching "the step of detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value (col. 15, ln. 1 to col. 16, ln. 14 and/or col. 14, ln. 4-55, the goal is to use 0 and 1 to represent for unvoiced and voice portions, respectively)." *Final Office Action*, page 4. However, *Yeldener* teaches that the low frequency portion of the signal spectrum contains a predominantly voiced signal, while thc high frequency portion of the spectrum contains predominantly the unvoiced portion of the speech signal. *Yeldener*, column 4, lines 34-37. In other words, in the *Yeldener* reference, low frequency equals voice, while high frequency equals no voice.

For example, *Yeldener* teaches that if Pv=1, the signal is purely voiced and only has harmonically related components; if Pv=0, the speech segment is purely unvoiced and can be modeled as a filtered noise. *Yeldener*, column 4, lines 44-48.

> The **unvoiced** portion of each time segment is reconstructed by selecting a codebook entry which comprises a **high pass filtered noise signal.** The codebook entries can be obtained from an inverse Fourier transform of the portion of the spectrum determined to be **unvoiced** by obtaining the spectrum of a **white noise signal** and then computing the

> inverse transform of the remaining signal in which **low frequency band components have been successively removed**. (emphasis added).

Yeldener, column 5, line 66 – column 6, line 7.

Hence, *Yeldener* teaches that the unvoiced portion of the segment is high frequency white noise in which the low frequency voiced signals have been removed.

In contrast, as amended, independent claim 1 of the present invention recites that the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voice. In other words, low frequency equals no voice, while high frequency equals voice as recited in claim 1. By way of example:

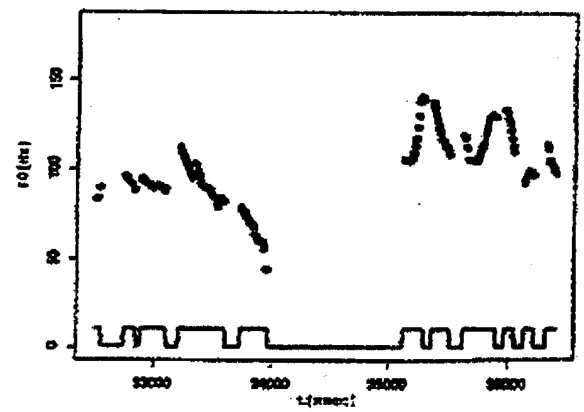

FIG. 4A

> Figure 4A shows a plot of F0 (in units of Hertz) over time (in units of milliseconds). The sampling rate is 10 milliseconds and thus the time distances between the plotted data is also 10 milliseconds so far as they are not interrupted by voiceless sections. Near its center, between 34000 and 35000 milliseconds, the plot comprises such a voiceless section that, in the underlying audio streams, separates two sentences. At the bottom of the diagram, an index function comprising values 1 (= ON) and 0 (= OFF) is depicted which is ON for voiced sections and OFF for voiceless sections of the audios stream.

Application, page 16, line 12 – page 17, line 6.

As Figure 4A of the present invention above illustrates, the voiced portion of the audio stream is high frequency and the unvoiced portion is low frequency. Consequently, claim 1 recites a claim limitation that directly contradicts the teachings of *Yeldener*. More specifically, claim 1 recites that low frequency equals no voice, while high frequency equals voice; whereas *Yeldener* teaches that low frequency equals voice, while high frequency equals no voice. As a result, *Yeldener* does not teach or suggest that the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voice as recited in claim 1. Thus, *Yeldener* does not teach or suggest the above recited claim limitation.

Therefore, because neither *Shriberg* nor *Yeldener* teach or suggest that the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voice as recited in claim 1, the combination of *Shriberg* and *Yeldener* cannot teach or suggest the above recited claim 1 limitation.

Furthermore, *Shriberg* fails to teach extracting and combining a plurality of prosodic features in a neighborhood of the candidate boundaries to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as recited in amended claim 1. *Shriberg* teaches extracting, computing, and normalizing prosodic features, *Shriberg*, page 130, section 1.4. But, *Shriberg* makes no reference to combining prosodic features. The focus is on the overall performance, and on analysis of which prosodic features proved most useful for each task. *Shriberg*, page 130, section 1.4. In determining which prosodic features are most useful, *Shriberg* analyzes each prosodic feature individually. *Shriberg* discusses and analyzes each prosodic feature class in individual sections. No section in *Shriberg* describes prosodic features in combination.

The Examiner cited section 2.1.4., on page 137, in order to demonstrate that *Shriberg* teaches the combination of a plurality of prosodic features. This cited section teaches that a feature selection algorithm was utilized to automatically reduce the initial candidate feature set to an optimal subset. *Shriberg*, page 137, section 2.1.4. Because the initial feature set in *Shriberg* contained over 100 features, the set is split into smaller

subsets. *Shriberg*, page 137, section 2.1.4. Features are grouped into broad feature classes based on the kinds of measurements involved, and the type of prosodic behavior they are designed to capture. *Shriberg*, page 131, section 2.1.2. In other words, the prosodic features are not extracted and combined in *Shriberg* in a neighborhood of the candidate boundaries to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as is recited in claim 1, but *Shriberg* merely places prosodic features into groups or classes depending upon the prosodic feature's behavior or measurements.

However, *Shriberg* does teach that for each task the results are examined from combining the prosodic information with language model information. *Shriberg*, page 130, section 1.4. In *Shriberg*, prosodic information is combined with language model information to evaluate overall performance. *Shriberg*, page 127, Abstract. But, combining prosodic and language models to evaluate performance is distinguishable from combining a plurality of prosodic features to determine boundaries for semantic and syntactic units in an audio stream as recited in amended claim 1. Therefore, *Shriberg* does not teach or suggest combining a plurality of extracted prosodic features in an audio segmentation process in order to determine semantic or syntactic units.

Again, *Yeldener* fails to cure the deficiencies of *Shriberg*. *Yeldener* teaches a method for providing encoding and decoding of speech signals using voicing probability determination. *Yeldener*, Abstract. More specifically, *Yeldener* teaches:

> …the input speech signal is represented as a sequence of time segments of predetermined length. For each input segment a determination is made as to detect the presence and estimate the frequency of the pitch F_0 of the speech signal within the time segment. Next, on the basis of the estimated pitch is determined the probability that the speech signal within the segment contains voiced speech patterns.

Yeldener, column 4, lines 25-32.

As the passage from *Yeldener* above teaches, only the prosodic feature of pitch is used to determine voiced speech pattern segments. In other words, only one prosodic feature is utilized in the method of *Yeldener*. Moreover, with regard to the *Yeldener* reference the Examiner stated:

> Regarding applicant's argument in the use of the Yeldener reference, Yeldener is only relied upon for the teaching of *detecting the changes of the fundamental frequency includes providing a threshold value for estimates of the fundamental frequency's voicedness and determining whether the voicedness of the fundamental frequency estimates are higher or lower than the threshold value*, as agreed by the applicant.

Final Office Action, pages 2 and 3.

As a result, *Yeldener* does not teach or suggest extracting and combining a plurality of prosodic features in a neighborhood of the candidate boundaries to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as recited in amended claim 1. Because neither *Shriberg* nor *Yeldener* teach or suggest the immediately preceding claim 1 limitation, the combination of *Shriberg* and *Yeldener* cannot teach or suggest this claim limitation.

Accordingly, the rejection of amended independent claim 1, which is representative of independent claims 11 and 12, as being unpatentable over *Shriberg* in view of *Yeldener* has been overcome. Therefore, amended independent claims 1, 11, and 12 are in condition for allowance. As a result, claims 3-6, 8, 10, and 13 are dependent claims depending on independent claims 1, 11, and 12, respectively. Consequently, claims 3-6, 8, 10, and 13 also are allowable, at least by virtue of their dependence on allowable claims. Furthermore, these dependent claims also contain additional features not taught by *Shriberg* and *Yeldener*.

For example, dependent claim 5 of the present invention reads as follows:

> 5. The method according to claim 4, wherein the environment is a time period between 500 and 4000 milliseconds.

With regard to claim 5, the Examiner stated:

> Regarding claims 4-6 and 10, Shriberg et al. further disclose a method for extracting at least one prosodic feature in an environment of the audio stream where the value of the index function is equal 0 *(section 2.1.1 on page 130 discusses feature extraction of both voice and unvoiced portions)*, **that the environment is a time period between 500 and 4000 milliseconds** ***(Section 2.1.1 on page 130)***, at least one prosodic feature is represented by the fundamental frequency *(Section 2.1.1, page 130)*, and

> a step of performing a prosodic feature classification based on a predetermined classification tree *(section 2.1.2 on page 131, grouping features)*. (emphasis added).

Final Office Action, pages 5 and 6.

Shriberg teaches that for each inter-word boundary prosodic features of the word immediately preceding and following the boundary are examined, or alternatively within a window of 20 frames or 200 milliseconds before and after the boundary. *Shriberg*, page 130, section 2.1.1. In other words, *Shriberg* utilizes a very specific time period to examine and extract prosodic features in the audio stream segmentation process. Thus, the only value "empirically optimized" for the method of *Shriberg* is 200 milliseconds. *Shriberg*, page 130, section 2.1.1.

In contrast, claim 5 recites that the environment for extracting a plurality of prosodic features from the candidate boundaries is a time period between 500 and 4000 milliseconds. Thus, *Shriberg* only teaches one time value of 200 milliseconds, whereas claim 5 recites a time internal of 500 to 4000 milliseconds. In addition, the 500 to 4000 millisecond range recited in claim 5 does not contain the 200 millisecond time value taught in *Shriberg*. Hence, *Shriberg* fails to teach or suggest the feature recited in dependent claim 5 of the present invention.

Accordingly, the rejection of claims 1, 3-6, 8, and 10-13 as being unpatentable over *Shriberg* in view of *Yeldener* has been overcome.

II. <u>35 U.S.C. § 103, Obviousness, Dependent Claims 9 and 14</u>

The Examiner has rejected dependent claims 9 and 14 under 35 U.S.C. § 103 as being unpatentable over *Shriberg* in view of *Yeldener*, as applied to claims 8 and 13 above, and further in view of Eryilmaz, U.S. Patent No. 5,867,574 ("*Eryilmaz*"). This rejection is respectfully traversed.

As shown in Section I above, *Shriberg* and *Yeldener* do not teach or suggest all the claim limitations recited in amended independent claims 1, 11, and 12. In particular, *Shriberg* and *Yeldener* do not teach or suggest that the voicedness of the fundamental frequency estimates lower than the threshold value equals no voice and the voicedness of the fundamental frequency estimates higher than the threshold value equals voice as

recited in claims 1, 11, and 12. Additionally, *Shriberg* and *Yeldener* do not teach or suggest that extracting and combining a plurality of prosodic features in a neighborhood of the candidate boundaries to determine boundaries for semantic and syntactic units depending only on the combined plurality of prosodic features as further recited in amended claims 1, 11, and 12. These recited features are also not taught or suggested in *Eryilmaz*.

Therefore, since *Shriberg*, *Yeldener*, and *Eryilmaz* do not teach or suggest the claim limitations recited above in amended claims 1, 11, and 12, then the combination of *Shriberg*, *Yeldener*, and *Eryilmaz* cannot teach or suggest these features. As a result, claims 9 and 14 of the present invention also are allowable at least by virtue of their dependence on allowable claims. Accordingly, the rejection of dependent claims 9 and 14 as being unpatentable over *Shriberg* in view of *Yeldener*, as applied to claims 8 and 13 above, and further in view of *Eryilmaz* has been overcome.

III. Conclusion

It is respectfully urged that the subject application is patentable over the cited references and is now in condition for allowance.

The Examiner is invited to call the undersigned at the below-listed telephone number if in the opinion of the Examiner such a telephone conference would expedite or aid the prosecution and examination of this application.

DATE: July 15, 2005

Respectfully submitted,

Peter B. Manzo

Peter B. Manzo
Reg. No. 54,700
Yee & Associates, P.C.
P.O. Box 802333
Dallas, TX 75380
(972) 385-8777
Attorney for Applicants